Journal of Theory and Practice of Engineering Science ISSN: 2790-1505
DOI: 10.53469/jtpes.2024.04(03).15

A Survey of Part-of-Speech Tagging

Linfeng Dai

School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China

Abstract: ***Part-of-Speech(POS) tagging, a fundamental task in natural language processing (NLP) that involves categorizing each word in a text into specific grammatical categories, is not only crucial for linguistic research but also serves as a prerequisite for more complex NLP applications such as syntactic analysis, entity recognition, and machine translation. This paper reveals the transition from the laborious process of manual annotation to the development of automated techniques, showcasing how the application of advanced deep learning (DL) and machine learning (ML) methods can enhance the efficiency and accuracy of POS tagging. Finally, the paper discusses the current challenges faced in POS tagging, along with corresponding solutions and potential future directions.***

Keywords: Part-of-Speech tagging, NLP, Machine Learning, Deep Learning.

1. INTRODUCTION

In the wide field of natural language processing (NLP), Part-of-Speech tagging (POS tagging) plays a crucial role. As a foundational task in NLP, Part-of-Speech tagging involves identifying the grammatical categories of each word in a text and assigning corresponding POS tags. The purpose of this review is to explore and summarize the development of POS tagging technology, ranging from early rule-based methods to the latest models based on deep learning[1]. Part-of-speech tagging is not only a fundamental tool for linguistic research but also a prerequisite for many complex NLP applications, such as syntactic analysis, entity recognition, and machine translation. Correct POS tagging is essential for understanding the precise meaning of words in specific contexts. Manual POS tagging is a tedious, time-consuming, and labor-intensive task[3]. Consequently, a growing interest in automated tagging processes is emerging in research. To achieve efficient POS tagging, researchers have begun to explore the potential of applying deep learning (DL) and machine learning (ML) techniques. These techniques are widely considered for their powerful capabilities in handling complex data and learning intricate patterns. Deep learning technologies, particularly neural networks, can improve the accuracy and efficiency of POS tagging through multi-layer representation learning and feature extraction. At the same time, machine learning methods, such as support vector machines and random forests, are also employed to extract linguistic features and perform effective classification. The aim of both deep learning and machine learning methods is to learn meaningful information from large training datasets[4-5]. The fusion and innovative application of these methods aim to enhance the performance of POS tagging tasks, thus enabling more accurate and efficient processing of natural language.

2. MACHINE LEARNING APPROACH FOR POS TAGGING

The Hidden Markov Model (HMM) is widely used in statistical machine learning[6-8]. The application of HMM in POS tagging primarily leverages its effectiveness in handling sequential data. In POS tagging tasks, each word represents an observed element in the sequence, while the Part-of-Speech of each word constitutes the hidden state. HMM predicts the most likely sequence of POS tags by calculating the probability of each tag's occurrence and the likelihood of a particular tag given a word. The strength of this model lies in its ability to capture contextual information, making it highly effective for understanding the syntactic structure of language.

In the method of using Support Vector Machines (SVM) for part-of-speech tagging, SVM is employed as a classifier to distinguish between different word classes. The process begins with preprocessing the text data and extracting features, which may include the word itself, its context, the POS tags of surrounding words, and more[9-11]. The SVM model is then trained using labeled training data, which consists of words and their corresponding POS tags. The SVM model learns to differentiate between various POS tags based on the given features. Once trained, the model can be applied to new text data to predict the POS tags of each word. In practical applications, SVM is considered an effective tool for POS tagging due to its strong performance and generalization capabilities in high-dimensional data.

In the method of using Conditional Random Fields (CRF) for POS tagging, CRF, as a sequence model, is used to tag the POS of each word in a sentence. The process includes: firstly, preprocessing the text data and extracting features such as the word itself, context information, and lexical structure. Then, these features and training data with POS tags are used to train the CRF model. The CRF model utilizes this information to understand the dependencies between words, thereby predicting the POS of each word more accurately. Once the model is trained, it can be applied to new text data to annotate the POS of words. CRFs are particularly effective in handling POS tagging tasks as they capture and utilize the dependencies in sequence data[11-15].

3. DEEP LEARNING APPROACH FOR POS TAGGING

Long Short-Term Memory networks (LSTMs) are an effective approach for part-of-speech tagging tasks, especially when dealing with sequential data like text. In this method, text is first preprocessed to extract features suitable for LSTM processing. The LSTM model's ability to remember long-term dependency information is invaluable for understanding the complex relationships between words in text. During the training phase, the model learns from training data annotated with POS tags, enabling it to predict the POS of each word in a sentence. Due to their capability to capture long-distance contextual information, LSTMs excel in POS tagging tasks, particularly when dealing with sentences that have complex structures and long-distance dependencies[18].

Using Recurrent Neural Networks (RNN) for POS tagging involves transforming text data into a sequence, enabling the RNN to process one word at a time. During the training phase, the RNN model learns from annotated datasets that contain words and their corresponding POS tags. A key feature of RNNs is their ability to process and utilize the contextual information in sequence data, as they consider the information of previous words while processing each word. This memory of context makes RNNs particularly effective in POS tagging tasks, enabling them to capture long-term dependencies in sentences and more accurately identify and tag the POS of each word. In practical applications, RNNs can analyze new text and predict the POS of each word, thereby supporting more complex natural language processing tasks[19].

Using Convolutional Neural Networks (CNN) for part-of-speech tagging involves first converting text data into sequences of word embedding vectors, as CNNs are adept at processing local features based on fixed window sizes. During the training phase, the CNN model learns from datasets annotated with POS tags, which include words and their corresponding POS tags. A key feature of CNNs is their use of convolutional layer filters to capture local features in the text, such as n-grams or the context around individual words. These features are extracted and combined through multiple convolutional and pooling layers, enabling the CNN to capture and utilize these local features for POS prediction. This sensitivity to local context makes CNNs particularly effective in POS tagging tasks, especially when dealing with texts that have dense semantic or syntactic features locally. In practical applications, a trained CNN model can effectively analyze new texts, accurately predicting the POS for each word, thereby supporting a broader range of natural language processing tasks[20].

4. 4.CHALLENGES AND FUTURE DIRECTIONS

4.1 Sample scarcity in some domains

One of the major challenges faced in POS tagging is sample scarcity, which is particularly evident in specific contexts or languages. For widely-used languages like English, there is a relative abundance of annotated corpora available for training and testing POS tagging models. However, for minority languages or specific dialects, there is a lack of sufficiently annotated datasets. Especially in specialized fields such as medicine and law, the terminology and linguistic structures used can differ significantly from general corpora, leading to fewer annotated samples in these areas. Additionally, the rise of the internet and social media has introduced a vast array of new vocabulary and informal expressions, often not included in traditional POS tagging training sets. The lack of diverse data in model training can lead to poor generalization, especially when dealing with vocabulary or text not present in the training set. For words that are rare or specific to certain fields, models might not accurately tag their POS, thereby affecting the quality of the entire text's analysis.

4.2 Future Technology Development Trends

In summary, POS tagging is one of the most important and fundamental tools for any other natural language processing task, such as information extraction, information retrieval, and machine translation. To address the issue of scarce training samples in specialized domains, as described earlier, it is recommended to use methods based on transfer learning. This involves using models pre-trained on large-scale corpora (such as BERT or GPT) and then fine-tuning them on small sample datasets. Alternatively, employing a multi-task learning framework, where the model simultaneously learns POS tagging and other related tasks (like syntactic parsing), can provide richer feature representations. This approach helps to leverage the knowledge gained from one task to improve performance on others, particularly useful in scenarios with limited domain-specific data.

5. CONCLUSION

In this paper, we have comprehensively explored the development and application of POS Tagging in the field of natural language processing (NLP). As one of the foundational tasks in NLP, POS tagging is crucial for understanding the grammatical category and precise meaning of each word in a text. This review detailed the evolution of POS tagging techniques, from early rule-based methods to modern approaches grounded in deep learning and machine learning. Special attention was paid to the challenges faced in POS tagging, especially the issue of sample scarcity in specific contexts, languages, or specialized domains. To address these challenges, we explored various strategies, including transfer learning and multi-task learning, which utilize pre-trained models and joint learning of different tasks to enhance the generalizability and performance of models.

In conclusion, POS tagging is not only a vital tool for understanding natural language but also a cornerstone supporting more complex NLP tasks, such as information extraction, information retrieval, and machine translation. Moving forward, as technology continues to advance and innovate, the methods and performance of POS tagging are expected to continually improve to meet the growing and diversifying demands of language processing.

REFERENCES

[1] Chiche A, Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning approaches[J]. Journal of Big Data, 2022, 9(1): 10.

[2] Kanakaraddi S G, Nandyal S S. Survey on parts of speech tagger techniques[C]//2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018: 1-6.

[3] Pisceldo F, Adriani M, Manurung R. Probabilistic part of speech tagging for Bahasa Indonesia[C]//Third international MALINDO workshop. 2009: 1-6.

[4] Alzubaidi L, Zhang J, Humaidi A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. Journal of big Data, 2021, 8: 1-74.

[5] Deshmukh R D, Kiwelekar A. Deep learning techniques for part of speech tagging by natural language processing[C]//2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2020: 76-81.

[6] Hasan F M, UzZaman N, Khan M. Comparison of different POS tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla[C]//Advances and innovations in systems, computing sciences and software engineering. Springer Netherlands, 2007: 121-126.

[7] Kuawat D, Jain V. POS tagging approaches: A comparison[J]. International Journal of Computer Applications, 2015, 118(6).

[8] Wicaksono A F, Purwarianti A. HMM based Part-of-Speech tagger for Bahasa Indonesia[C]//Fourth International MALINDO Workshop, Jakarta. 2010.

[9] Binulal G S, Goud P A, Soman K P. A SVM based approach to Telugu parts of speech tagging using SVMTool[J]. International Journal of Recent Trends in Engineering, 2009, 1(2): 183.

[10] Fernando S, Ranathunga S, Jayasena S, et al. Comprehensive part-of-speech tag set and svm based pos tagger for sinhala[C]//Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016). 2016: 173-182.

[11] Darwish K, Mubarak H, Abdelali A, et al. Arabic pos tagging: Don't abandon feature engineering just yet[C]//Proceedings of the third arabic natural language processing workshop. 2017: 130-137.

[12] Warjri S, Pakray P, Lyngdoh S A, et al. Part-of-Speech (POS) tagging using conditional random field (CRF) model for Khasi corpora[J]. International Journal of Speech Technology, 2021, 24(4): 853-864.

[13] Teller I, Eshkol I, Taalab S, et al. POS-tagging for oral texts with crf and category decomposition[J]. Research in Computing Science, 2010, 46: 79--90.

[14] Xu Z, Qian X, Zhang Y, et al. CRF-based hybrid model for word segmentation, NER and even POS tagging[C]//Proceedings of the sixth SIGHAN workshop on Chinese language processing. 2008.

[15] Warjri S, Pakray P, Lyngdoh S, et al. Adopting conditional random field (crf) for khasi part-of-speech tagging (kpost)[C]//Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India. Springer Singapore, 2021: 75-84.

[16] Horsmann T, Zesch T. Do LSTMs really work so well for POS tagging?–A replication study[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. 2017: 727-736.

[17] Horsmann T, Zesch T. Do LSTMs really work so well for POS tagging?–A replication study[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. 2017: 727-736.

[18] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[19] Shao Y, Hardmeier C, Tiedemann J, et al. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF[J]. arXiv preprint arXiv:1704.01314, 2017.

[20] Balwant M K. Bidirectional LSTM based on POS tags and CNN architecture for fake news detection[C]//2019 10th International conference on computing, communication and networking technologies (ICCCNT). IEEE, 2019: 1-6.